Am. J. Hum. Genet. 60:1174–1183, 1997

Genetic Relationships of Asians and Northern Europeans, Revealed by Y-Chromosomal DNA Analysis

Tatiana Zerjal,[1,*] Bumbein Dashnyam,[1,7] Arpita Pandya,[1] Manfred Kayser,[2] Lutz Roewer,[2] Fabrício R. Santos,[1] Wulf Schiefenhövel,[3] Neale Fretwell,[4] Mark A. Jobling,[4] Shinji Harihara,[5] Koji Shimizu,[6] Dashnyam Semjidmaa,[7] Antti Sajantila,[8] Pia Salo,[9] Michael H. Crawford,[10] Evgeny K. Ginter,[11] Oleg V. Evgrafov,[11] and Chris Tyler-Smith[1]

[1]Department of Biochemistry, University of Oxford, Oxford; [2]Institut für Gerichtliche Medizin, Humboldt-Universität zu Berlin, Berlin; [3]Forschungsstelle für Humanethologie, Max-Planck-Gesellschaft, Erling-Andechs, Germany; [4]Department of Genetics, University of Leicester, Leicester; [5]Department of Biological Sciences, University of Tokyo, Tokyo; [6]Department of Biology, Naruto University of Education, Naruto, Tokushima, Japan; [7]Department of Molecular Biology, Institute of Biotechnology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia; Departments of [8]Forensic Medicine and [9]Medical Genetics, University of Helsinki, Helsinki; [10]Laboratory of Biological Anthropology, Department of Anthropology, University of Kansas, Lawrence; and [11]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow

Summary

We have identified a new T→C transition on the human Y chromosome. C-allele chromosomes have been found only in a subset of the populations from Asia and northern Europe and reach their highest frequencies in Yakut, Buryats, and Finns. Examination of the microsatellite haplotypes of the C-allele chromosomes suggests that the mutation occurred recently in Asia. The Y chromosome thus provides both information about population relationships in Asia and evidence for a substantial paternal genetic contribution of Asians to northern European populations such as the Finns.

Introduction

Modern humans are thought to have originated in Africa <200,000 years ago and to have spread throughout the world during the past 100,000 years, but the details of these population movements are poorly understood (Lahr and Foley 1994). Analysis of protein and DNA polymorphisms can reveal the relationships between populations and thus allow past migrations to be identified (Cavalli-Sforza et al. 1994). Autosomal nuclear sequences and mtDNA have been used extensively for such purposes (Stoneking 1993), and Y-chromosomal markers are now starting to be used (Jobling and Tyler-Smith 1995). Modern Y chromosomes can be traced back to a common ancestor who is estimated to have lived ~188,000 years ago (Hammer 1995) or 37,000–49,000 years ago (Whitfield et al. 1995). The recent origin and high degree of geographical differentiation (Jobling and Tyler-Smith 1995) arising from its small effective-population size should make the Y chromosome particularly suitable for studying the spread of modern humans. Y markers have been used to investigate the origins of the Japanese (Hammer and Horai 1995) and to argue in favor of the "demic diffusion" model for the spread of Neolithic European farmers (Semino et al. 1996). We report here a new Y-chromosomal T→C transition, which appears to be restricted to a subset of the populations of Asia and Europe and which consequently provides an informative marker for the history of these areas.

Archaeological evidence suggests that parts of northern and central Asia probably have been inhabited continuously for the past 35,000 years, but humans could have entered northern Europe only after the Scandinavian ice sheet melted 8,000–15,000 years ago (Nunez 1987; Fagan 1995). Little is known about the origins of the Asian populations: there have been substantial recent population movements and acculturation, and the relationships of the modern populations are poorly understood (Forsyth 1992; Cavalli-Sforza et al. 1994). There are contradictory ideas about the origins of some of the northern European populations. Most Europeans speak Indo-European languages, but the Saami (also called "Lapps") in northern Scandinavia, the Finns, and the Estonians speak languages belonging to the quite different, Uralic (formerly called "Finno-Ugric") language group. This has led to the traditional view of their origin—that they have come from a "Finno-Ugric homeland" in central Asia (Sajantila and Pääbo 1995). Genetic analysis of classical markers (Cavalli-Sforza et al.

Received September 9, 1996; accepted for publication February 28, 1997.

Address for correspondence and reprints: Dr. Chris Tyler-Smith, CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom. E-mail: chris@bioch.ox.ac.uk

*Present affiliation: International Center for Genetic Engineering and Biotechnology, Trieste.

0002-9297/97/6005-0019$02.00

1994) and of mtDNA (Sajantila et al. 1995) confirms that the Saami are distinct from other European populations, but results for the Finns are less clear. Early studies provided evidence for some Asian admixture (Nevanlinna 1980), but more recent work has found that the Finns are indistinguishable from other European populations (Lahermo et al. 1996). It therefore has been suggested that the Finns are of European origin and originally spoke an Indo-European language but recently have adopted their present Uralic language (Sajantila and Pääbo 1995).

The new Y-chromosomal marker allows us to compare paternal lineages with the information provided by maternal lineages and language. It reveals that the European Uralic-speaking populations share with some central and northeastern Asians a Y-chromosome haplotype, providing genetic evidence for a substantial Asian paternal contribution to these northern European populations.

Material and Methods

Identification and Detection of the T→C Transition

The ends of the single-copy clone RBF5 (Tyler-Smith et al. 1993) in pTZ18R were sequenced by use of the Sequenase version 2.0 kit (Amersham) with the universal forward and reverse primers. The sequence information was used to design the primers R5D (5′-GTGAAGTAAGATATCAGATGG-3′) and R5I (5′-TGCAAGCTTAATTCATAGCAC-3′) and to amplify a 1.5-kb male-specific DNA fragment. PCR reactions were performed in a 25-μl volume containing 60 mM Tris-HCl pH 9.0, 15 mM $(NH_4)_2SO_4$, 2.0 mM $MgCl_2$, 100 μM dNTPs, 1.0 μM each primer, 1.25 units *Taq* polymerase (Promega), and 50–100 ng DNA; occasionally the volume was 50 μl. Thirty cycles of 94°C for 1 min, 61°C for 1 min, and 72°C for 1 min were used in an MJ Research PTC-200 thermal cycler. Twenty DNA samples, chosen to include at least one example of each haplotype defined by other Y markers (data not shown), were amplified. The product was digested with *AluI*+*RsaI* to produce fragments of 100–400 bp in size, which were analyzed by SSCP. Samples were denatured and run on a 36-cm-long 6% polyacrylamide gel (29:1, acrylamide:bis acrylamide) containing 5% glycerol in 0.5 × Tris-borate EDTA buffer at 230 V for 18 h. After electrophoresis, the gel was silver stained (Tegelström 1992), and a mobility difference was seen in one pair of bands in one sample (m227). The fragment was purified (QIAX II kit) from a male showing each mobility (m19 and m227) and was subcloned into the *SmaI* site of pBluescriptII KS+ (Stratagene). The 305-bp insert in three clones from each male was sequenced on both strands, and a single nucleotide difference was found between the two males. The primers Tat1 (5′-GACTCTGAGTGTAGACTTGTGA-3′) and Tat3 (5′-GAAGGTGCCGTAAAAGTGTGAA-3′) were designed to amplify a 112-bp fragment spanning the polymorphism. PCR conditions were as described above, except that the pH was 9.5 and the $MgCl_2$ concentration was 2.5 mM. The cycling protocol was 94°C for 1 min, 60°C for 1 min, and 72°C for 30 s.

Most population screening (table 1) was performed by use of the R5D and R5I primers and by digestion with either *Nla*III (New England Biolabs) or its isoschizomer *Hsp*92II (Promega), since the additional sites within the 1.5-kb fragment provided internal controls for monitoring the digestion. Digests were analyzed by electrophoresis in a 2.8% NuSieve, 0.8% agarose gel containing ethidium bromide in 0.5 × Tris-acetate EDTA (TAE) buffer. Degraded DNA samples were amplified by use of the Tat1 and Tat3 primers, and nested PCR with both sets of primers was used in some cases. The 112-bp fragment was analyzed by *Hsp*92II digestion and electrophoresis in a 4% NuSieve, 1% agarose gel containing ethidium bromide in 0.5 × TAE buffer. The sequence of most C-allele chromosomes was confirmed by digestion with *Mae*II (Boehringer).

Microsatellite Analysis

Seven Y-chromosomal microsatellite loci were selected from published papers (Roewer et al. 1992) and from the Genome Database (GDB), by use of the criteria of Y-chromosome specificity, a high level of polymorphism, and unequivocal identification of alleles. The following loci were used: *DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS393* (all tetranucleotide repeats), and *DYS392* (trinucleotide repeat). Primer sequences and PCR conditions were either as described elsewhere (Roewer et al. 1992) or as listed in the Genome Database, with the following modifications: PCR reactions were performed in a volume of 25 μl containing 20–100 ng DNA, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM each dNTP (Boehringer), 20 pmol each primer, with the forward primer 5′-labeled by indodicarbocyanine (CY5) phosphoramidite, and 1 unit *Taq* polymerase (Promega). PCR products were resolved and detected on 6% denaturing Hydrolink Long Ranger gels (AT Biochem) by use of the Automatic Laser Fluorescence express DNA sequencer (Pharmacia). Alleles are designated according to the fragment length determined by Fragment Manager Version 1.2 software (Pharmacia). Three additional trinucleotide microsatellites—*DYS425, DYS426,* and *DYF371*—were used as described elsewhere (Jobling et al. 1996).

Networks of the 10-element–microsatellite haplotypes were constructed separately for group 12 and C-allele chromosomes. Pairwise differences between hap-

Table 1

Frequency of T and C Alleles

Continent/Region and Population	No. of Alleles		
	T	C	Total
Africa:			
Kenyan	14	0	14
San	9	0	9
Algerian	27	0	27
Other	9	0	9
Europe:			
Italian	13	0	13
Albanian	10	0	10
Hungarian	39	0	39
Basque	26	0	26
German	71	0	71
United Kingdom	25	0	25
Icelandic	28	0	28
Norwegian	51	2	53
Finn	10	11	21
Saami	9	3	12
Estonian	10	9	19
Mari/Morkinsky	8	8	16
Mari/Gornomariysy	13	7	20
Mari/Orshansky	13	0	13
Mordva	7	2	9
Russian	17	3	20
Other	12	0	12
Asia:			
Indian	53	0	53
Sri Lankan	22	0	22
Buryat	47	64	111
Khalkh	46	1	47
Mjangad	1	1	2
Other Mongolian	14	0	14
Khalimag	0	1	1
Yakut	3	18	21
Altai	28	0	28
Keti	12	0	12
Evenki	25	0	25
Chinese	43	0	43
Japanese	163	1	164
Other	33	0	33
America:			
Amerindian, North	2	0	2
Amerindian, Central	3	0	3
Amerindian, South	22	0	22
Oceania:			
Trobriand Islands	63	0	63
Roro	13	0	13
Other	9	0	9
Total:	1,023	131	1,154
. . .:			
Chimpanzee	4	0	4
. . .:			
Orangutan	2	0	2

lotypes were tabulated, with each change of one microsatellite unit being counted as one difference. For *DYS389*, allowance was made for the smaller product *DYS389I* being contained within the larger product *DYS389II* (Cooper et al. 1996), so that, when both loci differed by a single repeat-unit change in the same direction, this was counted as only a single difference. The loci are designated "*DYS389a*" (smaller) and "*DYS389b*" (larger) in table 2, to indicate this. All haplotypes separated by a single difference were then linked by lines of unit length. Three haplotypes were separated by more than one difference and were linked by lines two or three units long.

DNA Samples

The DNA samples analyzed included subsets of the 91 described elsewhere (Mathias et al. 1994) and of the Y Chromosome Consortium collection (Hammer and Ellis 1995), samples collected by the authors, and additional samples provided by Adolfo López de Munain (Basques), Doudja Nafa (Algerians), Sveinn Guddmundsson (Icelanders), Gaute Brede and Hans Prydz (Norwegians), Giorgio Graziosi (Italians), Vladimir Osakovsky (Yakut), and Katalin Rajczy (Hungarians). Male and female chimpanzee and orangutan samples were from the National Institute of General Medical Sciences' Human Genetic Mutant Cell Repository (Camden, NJ), the American Type Culture Collection (Bethesda), and the Genetics Laboratory, Oxford.

Results

A New T→C Transition

The polymorphic T/C nucleotide lies within the single-copy locus RBF5 (Tyler-Smith et al. 1993) in proximal Yq. Amplification of male DNA with the primers Tat1 and Tat3 produces a 112-bp product containing the sequence CATGT (T allele) or CACGT (C allele) (see fig. 1*a* and *b*). By a fortunate coincidence, the T allele is cleaved by either *Hsp*92II or its isoschizomer *Nla*III and is resistant to *Mae*II, whereas the C allele is resistant to *Hsp*92II and is cleaved by *Mae*II, allowing the sequence of each allele to be determined unambiguously by restriction-enzyme digestion (fig. 1*c*). Alternatively, the external primers R5I and R5D can be used to amplify a ~1.5-kb fragment showing a more complex digestion pattern (fig. 1*c*), or the two sets can be used for nested PCR. Chimpanzees and orangutans contained the T allele, suggesting that T is ancestral.

Geographical Distribution

The allele present in each of 1,154 males was determined (table 1). Africans, southern Europeans, southern Asians, Oceanics, and Americans contained only the T allele, whereas both the T allele and the C allele were found in several populations from Asia and northern Europe. The frequency of the C allele ranged from ~1% in the Japanese to ~86% in the Yakut (Siberia) (fig. 2). It also was present in 2% of the Khalkhs (Mongolia)

analyzed, 4% of the Norwegians, 15% of the Russians, 25% of the Saami, 31% of the Mari (Uralic-language speakers in Russia), 47% of the Estonians, 52% of the Finns, and 58% of the Buryats (Mongolia).

The T→C transition has occurred on a Y haplotypic background designated "group 12," defined by the LLY22g/*Hin*dIII polymorphism (data not shown). Group 12 chromosomes have the LLY22g/*Hin*dIII polymorphism but the Tat T allele, whereas C-allele chromosomes have the LLY22g/*Hin*dIII polymorphism and the Tat C allele. Group 12 chromosomes have a wide geographical distribution and are found in southern and eastern Asia, but they have not been detected in Africa (Jobling et al. 1996; data not shown [the C-allele chromosomes are a subset of the Y chromosomes carrying a 50f2/C deletion]). The mutation thus may have occurred after the entry of modern humans into Asia but before their dispersal within Asia and Europe.

Origin of the Mutation

In order to obtain a better understanding of the history of this mutation, we have determined the haplotypes of a set of 60 C-allele chromosomes, at 10 microsatellite loci (table 2). The samples used were chosen on the basis of DNA availability, except that only 22 of the 64 available Buryat DNAs were used. Three widely variable microsatellites—*DYS426, DYF371,* and *DYS425* (Fretwell 1997)—show no variation, suggesting that the T→C transition only occurred once, but the other seven loci reveal 21 compound haplotypes. These haplotypes can be linked readily to form a network where most of the links require just one single-step–slippage mutation (fig. 3). Each haplotype generally is confined to a single population, showing that there has not been extensive recent admixture between the populations. Where samples of >10 chromosomes are available, each population contains multiple haplotypes: six haplotypes are found among 22 Buryats, six among 18 Yakut, and five among 11 Finns. Thus the diversity of C-allele chromosomes within each population is similar. Haplotypes from the same population generally cluster in the network (fig. 3). Furthermore, the network can readily be superimposed on the map of the world (fig. 4), revealing a striking correlation between Y haplotype and geographical location.

We have investigated whether it is possible to root the network and thus identify where the T→C transition occurred. A similar microsatellite analysis was performed on chromosomes from an outgroup, group 12, consisting of two chromosomes from Mongolia, four from China, and one from India, to define six haplotypes, T1–T6 (table 2). Two possible links between the two networks were considered: through haplotypes 18 and T6 and through haplotypes 2 and T2, T3, and T5. The 2/T2-T3-T5 link (fig. 3) is more probable, both because it results in a smaller mean number of differences between the C-allele and group 12 chromosomes and because it takes place between individuals who are geographically close (Mongolia/China) rather than far apart (Finland/India). If correct, this would identify haplotype 2 as the root and would identify Mongolia as a candidate for the location of the T→C transition (fig. 4).

We have used two approaches to investigate whether it is possible to determine when the T→C mutation occurred. If the variants were selectively neutral, and if microsatellite mutation rate and population history were known, we could estimate how long the present microsatellite variation had taken to accumulate. The mutation rates of Y microsatellite loci are not known, but a figure of 2.1×10^{-3} is available for autosomal tetranucleotide repeats (Weber and Wong 1993). In the first approach (Goldstein et al. 1996), it is assumed that the T→C mutation occurred only once, so that the microsatellite variance associated with the C-allele chromosomes was initially zero, and that it has increased with time. If equilibrium has not been reached, the present variance can be used to estimate the time required. The mean variance observed for the 10 microsatellite loci is 0.22. The 95% confidence limits for the variance expected at equilibrium, for an effective population size of 4,500 and a mutation rate of 2.1×10^{-3}, are 4.2–18.4. Equilibrium therefore has not been reached, and the observed variance corresponds to ~108 generations (95% confidence limits 92–128 generations), or ~2,200 years at 20 years/generation. In the second approach, based on that used to date the ΔF508 cystic fibrosis mutation (Bertranpetit and Calafell 1996), the mean number of steps, in the network, between the root (assumed to be haplotype 2) and each of the 60 chromosomes was counted. For the same mutation rate and generation time, the observed figure of 0.31 mutations/locus/chromosome corresponds to ~3,000 years. If only the six tetranucleotide Y microsatellites were used in these calculations, on the basis that the measurement of mutation rate is valid only for tetranucleotide loci, the first approach would give a figure of ~2,400 years, and the second would give a figure of 4,000 years.

Discussion

Our results illustrate the information that can be obtained from Y-chromosomal studies by use of haplotypes containing both slowly evolving base-substitution markers and more rapidly evolving microsatellites. By combining the Y data with previously published work, we can compare paternal lineages with language and the genetic information from autosomal and maternal lineages and thus can understand better the history of the populations of parts of Asia and Europe.

The T→C transition is likely to have occurred only

Table 2

Microsatellite Haplotypes of C-Allele and Group 12 Chromosomes

Chromosome Type, Haplotype, and Sample[a]	Population[b]	*DYS426* 94	*DYF371* 198, 201	*DYS425* 201	*DYS19* 190	*DYS389a* 255[c]	*DYS389b* 114[c]	*DYS390* 211	*DYS391* 287	*DYS392* 257	*DYS393* 128
C allele:											
1:											
m209	Buryat	—	—, —	—	—	—	—	—	↓	—	—
m227	Buryat	—	—, —	—	—	—	—	—	↓	—	—
D7	Buryat/Khugduud	—	—, —	—	—	—	—	—	↓	—	—
D8	Buryat/Khargana	—	—, —	—	—	—	—	—	↓	—	—
D15	Buryat/Tsagaanguud	—	—, —	—	—	—	—	—	↓	—	—
D45	Buryat/Sharaid	—	—, —	—	—	—	—	—	↓	—	—
D48	Buryat/Khargana	—	—, —	—	—	—	—	—	↓	—	—
D63	Buryat/Khalbin	—	—, —	—	—	—	—	—	↓	—	—
D92	Buryat/Sharaid	—	—, —	—	—	—	—	—	↓	—	—
D110	Buryat/Khugduud	—	—, —	—	—	—	—	—	↓	—	—
D142	Buryat/Bodonguud	—	—, —	—	—	—	—	—	↓	—	—
D151	Buryat/Khargana	—	—, —	—	—	—	—	—	↓	—	—
D159	Buryat/Bodonguud	—	—, —	—	—	—	—	—	↓	—	—
D27	Khalimag	—	—, —	—	—	—	—	—	↓	—	—
2:											
D55	Buryat/Khugduud	—	—, —	—	—	—	—	—	↓	—	↓
D152	Buryat/Khugduud	—	—, —	—	—	—	—	—	↓	—	↓
3:											
D130	Buryat/Khugduud	—	—, —	—	—	—	—	↑	↓	—	—
4:											
D114	Buryat/Sharaid	—	—, —	—	—	—	—	—	↓	—	↑
5:											
D13	Buryat/Sharaid	—	—, —	—	—	—	—	—	—	—	↑
6:											
m253	Khalkh	—	—, —	—	—	—	—	—	—	—	—
m274	Mjangad	—	—, —	—	—	—	—	—	—	—	—
R1247	Russian	—	—, —	—	—	—	—	—	—	—	—
m295	Norwegian	—	—, —	—	—	—	—	—	—	—	—
7:											
To121	Japanese	—	—, —	—	—	—	—	↓	↓	—	↓
8:											
m203	Buryat	—	—, —	—	—	↓	—	—	↓	—	—
D23	Buryat/Khargana	—	—, —	—	—	↓	—	—	↓	—	—
D49	Buryat/Sharaid	—	—, —	—	—	↓	—	—	↓	—	—
D125	Buryat/Sharaid	—	—, —	—	—	↓	—	—	↓	—	—
9:											
R1122	Russian	—	—, —	—	—	↓	↑	—	—	—	—
10:											
R1322	Russian	—	—, —	—	—	—	↑	—	—	—	—
11:											
JK3150	Yakut	—	—, —	—	—	—	↑	—	—	↑	—
JK3151	Yakut	—	—, —	—	—	—	↑	—	—	↑	—
ЯМ7	Yakut	—	—, —	—	—	—	↑	—	—	↑	—
12:											
ЯМ13	Yakut	—	—, —	—	—	—	↑	—	↓	↑↑	—
13:											
JK3146	Yakut	—	—, —	—	—	—	↑	—	—	↑↑	—
JK3152	Yakut	—	—, —	—	—	—	↑	—	—	↑↑	—
ЯК17	Yakut	—	—, —	—	—	—	↑	—	—	↑↑	—
ЯК18	Yakut	—	—, —	—	—	—	↑	—	—	↑↑	—
14:											
ЯК21	Yakut	—	—, —	—	—	—	↑↑	—	↓	↑↑	—
15:											
JK3149	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ1	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ3	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ6	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ11	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ18	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ19	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—
ЯМ24	Yakut	—	—, —	—	—	—	↑↑	—	—	↑↑	—

(*continued*)

Table 2 (continued)

Chromosome Type, Haplotype, and Sample[a]	Population[b]	*DYS426* 94	*DYF371* 198, 201	*DYS425* 201	*DYS19* 190	*DYS389a* 255[c]	*DYS389b* 114[c]	*DYS390* 211	*DYS391* 287	*DYS392* 257	*DYS393* 128
16:											
ЯК8	Yakut	—	—, —	—	—	—	↑↑↑	—	—	↑↑	—
17:											
LGL5144	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5190	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5209	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5236	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5246	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5254	Finn	—	—, —	—	—	—	—	↑	—	—	—
LGL5298	Finn	—	—, —	—	—	—	—	↑	—	—	—
N251	Norwegian	—	—, —	—	—	—	—	↑	—	—	—
18:											
LGL5198	Finn	—	—, —	—	—	—	—	↑	—	—	↓
19:											
m121	Finn	—	—, —	—	—	—	↓	↑	—	—	↓↓
20:											
LGL5293	Finn	—	—, —	—	—	—	—	↑↑	↑	—	—
21:											
LGL5191	Finn	—	—, —	—	↑	—	—	—	↓	—	—
Group12:											
T1:											
m205	Khalkh	—	—, —	—	—	↓	↓	—	↓	—	↓
m243	Khalkh	—	—, —	—	—	↓	↓	—	↓	—	↓
T2:											
m297	Chinese/Han	—	—, —	—	—	—	↑	—	↓	—	↓
T3:											
m300	Chinese/Han	—	—, —	—	—	—	—	—	↓	↓	↓
T4:											
m307	Chinese/Han	—	—, —	—	—	—	↑	—	↓	↑	↓
T5:											
m326	Chinese/Han	—	—, —	—	—	—	↓	—	↓	—	↓
T6:											
m462	Indian	—	—, —	—	—	—	—	↑	—	—	↓

NOTE.—A dash (—) denotes same size; ↑ = one unit larger; ↑↑ = two units larger; ↑↑↑ =three units larger; ↓ = one unit smaller; and ↓↓ = two units smaller.

[a] Code name.

[b] As listed in table 1; for Buryats, tribes are also given.

[c] As described in Material and Methods.

once: point mutations on the Y chromosome are rare (Hammer 1995; Jobling and Tyler-Smith 1995; Whitfield et al. 1995), and haplotyping the C-allele chromosomes by use of 18 other markers representing rare mutational events suggests a single origin (data not shown). C-allele chromosomes also show a very limited range of microsatellite haplotypes that form a simple network (fig. 3). In comparison, *DYS19* shows great variability in different populations (Santos et al. 1995), and just four of the loci used here (*DYS19, DYS389I, DYS389II* and *DYS390*) revealed 77 different haplotypes in a sample of 159 Dutch and German males that formed a much more complex network (Roewer et al. 1996). Similarly, seven of the loci (*DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392,* and *DYS393*) produced 27 different haplotypes in a sample of 38 Mongolians (Khalkhs, carrying the T allele) and produced 34 haplotypes in 36 Chinese (Han, carrying the T allele), emphasizing the high level of polymorphism of these Y-chromosomal microsatellites (data not shown).

The mutation probably occurred in Asia, since group 12 chromosomes have been found only in Asia, and the root for the haplotype network most probably was among the ancestors of the Buryats who now reside in Mongolia. Its most likely location thus was in the Mongolia/China area. Its date is uncertain: the estimates obtained were ~2,000–4,000 years ago. It is difficult to provide confidence limits for these estimates. Two major sources of error are the mutation rate, where a measurement specific to the Y loci used is required and allele length should also be taken into account, and the generation time: we have used 20 years, but longer or shorter times would produce proportionately earlier or later dates. Allele loss due to drift should be taken into account also. Because of these uncertainties, we probably could not exclude a date of, for example, 10,000

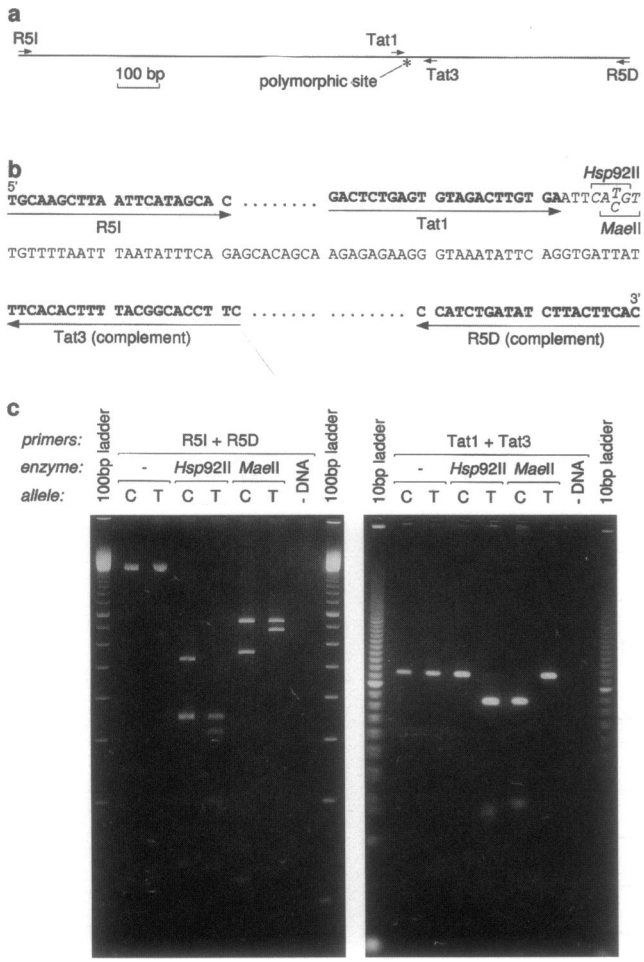

Figure 1 *a,* Structure of RBF5 locus, showing locations of primers and polymorphic site. *b,* Sequence of regions containing primers and polymorphic site. Note that the sequences shown here are complementary to the Tat3 and R5D primers. *c,* Scoring of C and T alleles. C-allele DNA (m227) or T-allele DNA (m19) was amplified with the primers shown at the top, was digested with the enzyme indicated, and was analyzed by gel electrophoresis.

plained either by genetic drift, which may have been extensive in small isolated populations in Siberia, or by language replacement in some populations. Uralic languages are divided into Finno-Ugric and Samoyedic; the Finno-Ugric branch is divided further, into Finno-Permic (including Mari, Estonian, Finnish, Saami, and Mordvinic) and Ugric (including Hungarian) (Grimes 1996). Within the Finno-Permic branch there is a good correlation between Y chromosome and language: if the Mari are grouped together, moderate frequencies of the C allele (22%–52%) are found among all Finno-Permic–division speakers examined. A few C-allele chromosomes are found among Japanese, Russian, and Norwegian speakers, perhaps reflecting admixture from neighboring Altaic- or Uralic-speaking populations.

In Asia, little is known about the genetic relationships of the Yakut to other populations (Cavalli-Sforza et al. 1994), but previous work using protein polymorphisms has shown a close relationship between Buryats and Mongolians (largely Khalkhs; Novoradovsky et al. 1993); similarly, studies of mtDNA have detected minimal levels of genetic substructuring between different populations in Mongolia (Kolman et al. 1996). This contrasts with the striking difference, in frequency of the Y-chromosomal C allele, between Buryats and Khalkhs that has been noted above. Within the Buryats, there is even a significant frequency difference between those from Dashbalbar sum, where 64% (46/72) have the C allele, and those from Tsagaan-Ovoo sum, where 42% (15/36) have the C allele ($P < .05$). The lower frequency of C-allele chromosomes in the Tsagaan-Ovoo sum population may be due to admixture during migrations, in this century, to China and then back to Mongolia. No significant differences are seen between different Buryat tribes; such differences would not be expected, since the traditional custom of marriage between Buryats from different tribes would lead to similar

years ago for the mutation. Nevertheless, all of these dates are recent, and the present distribution of the mutation, stretching from Japan in the east to Norway in the west, implies high levels of male-mediated gene flow in Asia and Europe during the past few thousand years.

Most C-allele chromosomes are found among speakers of two language families: Altaic and Uralic. Altaic languages are classified into three main groups: Mongolian (including Buryat and Khalkh), Tungus (including Evenki), and Turkic (including Altai and Yakut) (Grimes 1996). High frequencies of C-allele chromosomes were found in the Yakut (86%) and Buryats (58%), whereas the frequencies in the other populations were either low (2% in Khalkhs) or zero (Evenki and Altai). The absence of a simple correspondence between Y chromosome and language classification within this family could be ex-

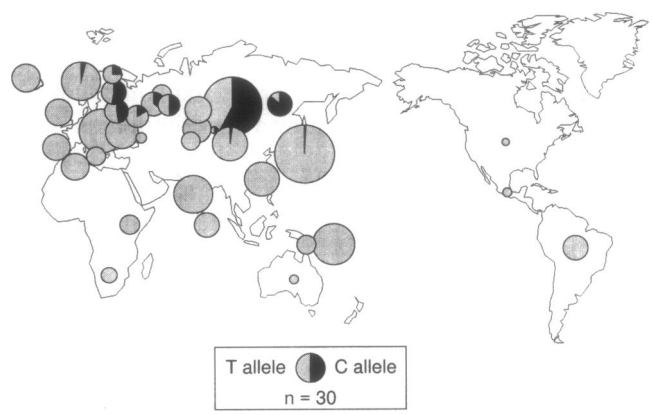

Figure 2 Geographical distribution of the T allele (*gray*) and C allele (*black*). The area of each circle is proportional to the number of individuals tested.

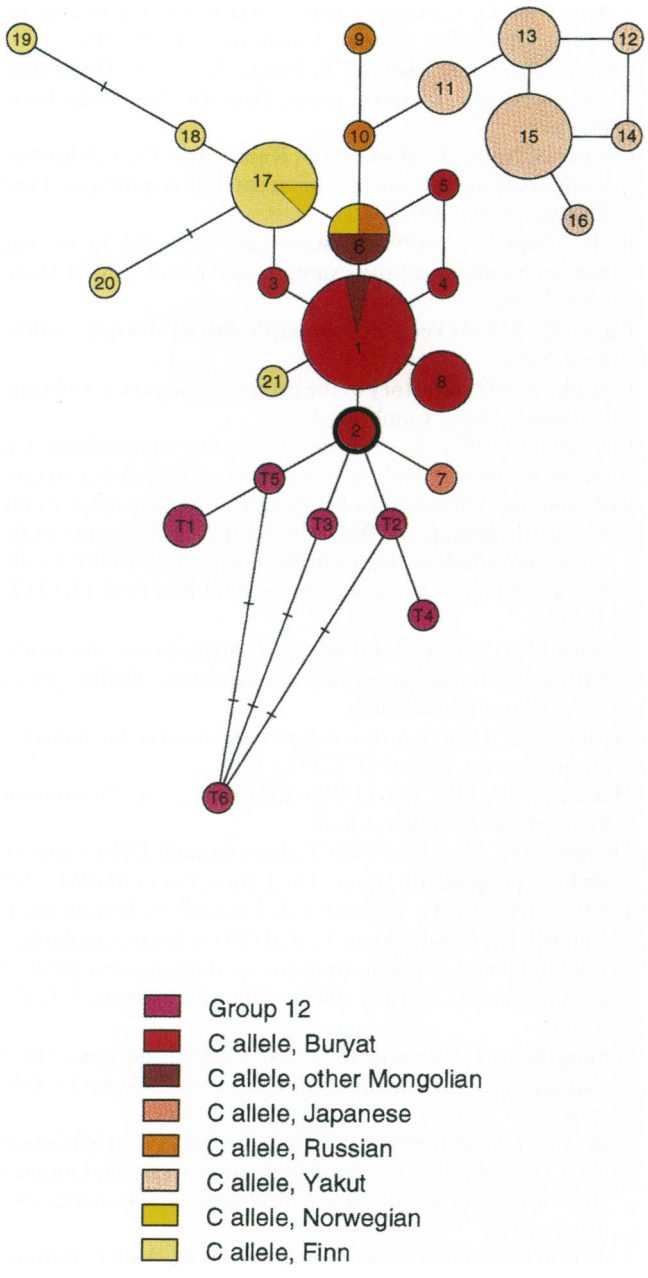

Figure 3 Network of C-allele and group 12 microsatellite haplotypes. Haplotypes are defined in table 2. The area of each circle is proportional to the frequency of the haplotype, and the color indicates the population of origin. Haplotype 2 (*thick edge*) is the most likely root. Lines represent one microsatellite mutation, except for the longer ones with bars, which represent two or three mutations.

frequencies in each tribe. C-allele chromosomes extend into Japan, possibly indicating a minor and relatively recent contribution from the region of Mongolia to the Japanese population.

In Europe, more historical and genetic information is available. Classical markers and mtDNA reveal little differentiation between central and northern populations, except for the Saami (Cavalli-Sforza et al. 1994; Sajantila et al. 1995). There is again a striking contrast with the data from the Y chromosome, where the C allele is present in 52% of the Finns, 47% of the Estonians, and 31% of the Mari but not at all in several other populations, such as the Icelanders, Germans, or British. Since the C-allele chromosomes probably have a single origin in Asia, the high frequency in the northern European populations might be explained formally by (1) differential migration of males from Asia to Europe; (2) differential gene flow, with either European autosomal and mtDNA sequences—but not Y-chromosomal DNA sequences—entering a population of Asian origin or Asian Y chromosomes entering a population of European origin; or (3) either genetic drift in small founder populations containing both European and Asian elements or subsequent bottlenecks. These explanations are not mutually exclusive. The presence of the highest European frequencies of the C allele in the Mari, Mordva, Estonians, Finns, and Saami, all Uralic speakers, suggests that the chromosomes have been carried westward by migrations of Uralic-speaking populations that extended as far as Finland. In this case, the Finns could have retained their original language and Y chromosomes but could have replaced most of their mtDNA and autosomal DNA by European sequences (Sajantila and Pääbo 1995; Lahermo et al. 1996). There is thought to have been a bottleneck in the Finnish population ~2,000–2,500 years ago (de la Chapelle 1993), but the relationship of this bottleneck to the postulated population-genetic changes is unclear. C-allele chromosomes are present at lower frequency in the Russians (15%) and are widespread: the three chromosomes analyzed come from the Ukraine, the Tambov region, and Turkmenistan and may represent admixture with neighboring Uralic speakers. The available haplotype information (fig. 3) suggests the possibility of a link between most of the Finnish C-allele chromosomes and the Bur-

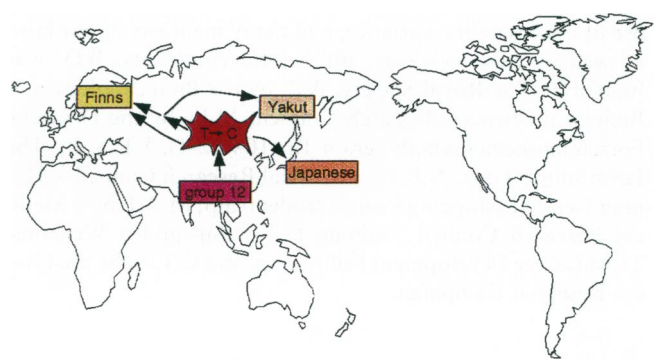

Figure 4 Origin and spread of the C-allele chromosomes. A simplified form of the haplotype network shown in figure 3 has been superimposed on a map of the world.

yat chromosomes, through haplotype 6, the only haplotype that is geographically widespread and extends across Asia and Europe, and thus this haplotype information is consistent with spread by Uralic speakers.

The C-allele chromosome is likely to have been present in many of the populations ancestral to the modern inhabitants of central and northern Asia and Europe. If these populations were at times small and isolated, drift could have led to high frequencies in some and to low (or zero) frequencies in others. The language spoken by any mixed population also could have been subject to "drift." Thus a complex pattern of association or nonassociation between Y haplotype and language could have been established. Subsequent population expansions and migrations with limited admixture would have retained many features of the pattern and could produce the striking geographical distribution that we now see. Analysis of ancient DNA potentially can link genetic changes to specific historical populations, and the T/C polymorphism would be well suited to such a study.

In conclusion, this work provides an example of the way in which a combination of point-mutation and microsatellite polymorphisms can be used to construct haplotypes and to trace Y lineages in great detail—and to link them to the geographical distribution of populations. Additional unique mutations and more reliable methods of dating, using microsatellites, are now needed. With these, it will be possible to connect all human Y lineages into a single calibrated network that summarizes human male history.

Acknowledgments

We thank all the original DNA donors for making this study possible; Kamal Bagai, Gaute Brede, Upen De Zylva, John Edwards, Nathan Ellis, Giorgio Graziosi, Sveinn Guddmundsson, Michael Hammer, Raoul Heller, Adolfo López de Munain, Doudja Nafa, V. L. Osakovsky, Hans Prydz, Katalin Rajczy, Lalji Singh, and Yuuichi Tanabe for help in obtaining samples; Albert de la Chapelle for useful discussions about the origins of the Finns; David Goldstein for advice on dating by use of microsatellite variance; and many members of our labs, as well as two reviewers, for helpful comments. B.D. was funded by The Royal Society, A.P. by the Biotechnology and Biological Sciences Research Council, M.K. by the Deutsche Forschungsgemeinschaft (grant Ro 1040/2-1), F.R.S. by The Leverhulme Trust, N.F. by a Medical Research Council—Human Genome Mapping Project studentship, M.A.J. by a Medical Research Council Training Fellowship and a Wellcome Trust Career Development Fellowship, and C.T.-S. by the Cancer Research Campaign.

References

Bertranpetit J, Calafell F (1996) Genetic and geographic variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds) Variation in the human genome. John Wiley & Sons, Chichester, pp 97–118

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

de la Chapelle A (1993) Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 30:857–865

Fagan BM (1995) People of the earth, 8th ed. Harper Collins, New York

Forsyth J (1992) A history of the peoples of Siberia. Cambridge University Press, Cambridge

Fretwell N (1997) A search for Y-specific minisatellites and microsatellites. PhD thesis, University of Leicester, Leicester

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13:1213–1218

Grimes BF (1996) (ed) Ethnologue: languages of the world, 13th ed. Summer Institute of Linguistics, Dallas (http://www.sil.org/ethnologue/)

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hammer MF, Ellis NA (1995) Appendix 1. Y Chromosom Consortium Newslett 2:8–9

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767–1775

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11:449–456

Kolman CJ, Sambuughin N, Bermingham E (1996) Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics 142: 1321–1334

Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L, Savontaus M-L (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am J Hum Genet 58:1309–1322

Lahr MM, Foley R (1994) Multiple dispersals and modern human origins. Evol Anthropol 3:48–60

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Nevanlinna HR (1980) Genetic markers in Finland. Haematologia 13:65–74

Novoradovsky AG, Spitsyn VA, Duggirala R, Crawford MH (1993) Population genetics and structure of Buryats from the Lake Baikal region of Siberia. Hum Biol 65:689–710

Nunez MG (1987) A model for the early settlement of Finland. Fennoscandia Archaeol 4:3–18

Roewer L, Arnemann J, Spurr NK, Grzeschik K-H, Epplen JT

(1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89:389–394

Roewer L, Kayser M, Dieltjes P, Nagy M, Bakker E, Krawczak M, de Knijff P (1996) Analysis of molecular variance (AMOVA) of Y-chromosome–specific microsatellites in two closely related human populations. Hum Mol Genet 5: 1029–1033

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus M-L, Aula P, et al (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5:42–52

Sajantila A, Pääbo S (1995) Language replacement in Scandinavia. Nat Genet 11:359–360

Santos FR, Gerelsaikhan T, Munkhtuja B, Oyunsuren T, Epplen JT, Pena SDJ (1995) Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism *DYS19*. Hum Genet 97:309–313

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS (1996) A view of the Neolithic demic diffusion in Europe through two Y chromosome-specific markers. Am J Hum Genet 59:964–968

Stoneking M (1993) DNA and recent human evolution. Evol Anthropol 2:60–73

Tegelström H (1992) Detection of mitochondrial DNA fragments. In: Hoelzel AR (ed) Molecular genetic analysis of populations. IRL Press, Oxford, pp 89–113

Tyler-Smith C, Oakey RJ, Larin Z, Fisher RB, Crocker M, Affara NA, Ferguson-Smith MA, et al (1993) Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. Nat Genet 5:368–375

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379–380